# SemanticTopoLoop: Semantic Loop Closure With 3D Topological Graph Based on Quadric-Level Object Map

Zhenzhong Cao<sup>1</sup>, Jingtai Liu<sup>1\*</sup>

Abstract-Loop closure, as one of the crucial components in SLAM, plays an essential role in correcting the accumulated errors. Traditional appearance-based methods, such as bag-ofwords models, are often limited by local 2D features and the volume of training data, making them less versatile and robust in real-world scenarios, leading to missed detections or false positives detections in loop closure. To address these issues, we first propose a semantic loop closure method based on quadriclevel object map topology, which represents scenes through the topological graph of quadric-level objects and achieves accurate loop closure at a wide field of view by comparing differences in the topological graphs. Next, in order to solve the data association problem between frame and map in loop closure, we propose a object-level data association method based on multilevel verification, which can associate 2D semantic features of current frame with 3D objects landmarks of map. Finally, we integrate these two methods into a complete object-aware SLAM system. Qualitative experiments and ablation studies demonstrate the effectiveness and robustness of the proposed object-level data association algorithm. Quantitative experiments show that our semantic loop closure method outperforms existing state-of-theart methods in terms of precision, recall and localization accuracy metrics.

## I. INTRODUCTION

Loop closure is indispensable for service robots operating in indoor environments, as they often need to navigate repetitive routes during their long-term operation. Traditional visual SLAM algorithms typically treat loop closure as a scene recognition problem, using low-level 2D features (such as SIFT [1], ORB [2], etc.) extracted from images for scene recognition and matching, such as the ORB-SLAM series algorithms [3]–[5]. and Vins-Mono [6], based on the DBow2 model [7]. With the advancement of deep learning (e.g., YOLO [8], Mask RCNN [9]), extracting semantic information from images has become more convenient, leading to the emergence of numerous loop closure methods based on semantic information, which not only can provide richer information for scene recognition, utilizing higher-level scene details, but also can exhibit greater robustness during scene matching, accommodating significant changes in the scene's perspective.

Due to the abundance of object information in indoor environments, object construction has become mainstream in SLAM processes, leading to the emergence of several excellent object-level SLAM algorithms. Consequently, object-level semantic data association methods and semantic loop closure methods based on object information have also emerged. However, current object-level semantic data association suffers from issues of low robustness and accuracy in complex scenarios such as false positives and false negatives of object detection network, occlusions and other abnormal conditions. Existing semantic loop closure methods based on object information are prone to false loop detections in repetitive scenes and missed loop detections in scenes with significant perspective changes.

In this paper, to address these challenges, we first introduce a multi-level verification based object-level data association method (abbreviated as MLV-ODA method), which resolves the challenging problem of data association between detetctions of current frame and quadric-level object landmarks of map, particularly in complex scenes where issues like false positives, false negatives, and occlusions are prevalent. This method achieves efficient and accurate data association results and provides prior information for scene representation and matching. Then, we present a quadric-level object map topology based semantic loop closure method (abbreviated as **OLT-SLC** method), addressing the issues of detecting false positive loop closures under significant viewpoint changes and improving the robustness of loop closure detection. This method enhances both the accuracy and recall of loop closure detection while providing precise loop closure candidates for subsequent global pose correction. Next, we embed the proposed MLV-ODA method and QLT-SLC method into the Object-Aware SLAM system that we proposed in [10], forming a complete semantic SLAM system that jointly maintains posepoint-object map database (abbreviated as **PPO-MD**). Finally, We conduct reasonable qualitative experiments, quantitative experiments, and ablation studies to validate the effectiveness and robustness of the proposed MLV-ODA and QLT-SLC methods in a variety of complex indoor scenes.

## The main contributions of this work are as follows:

- MLV-ODA method is introduced to reduce the time and space complexity of data association, indirectly promoting the accuracy and completeness of object construction in the scene.
- QLT-SLC method is presented to improve the precision and recall rate of loop closure, as well as enhance the system's localization accuracy.
- The proposed MLV-ODA method and QLT-SLC method are embed into the Object-Aware SLAM system, which jointly maintain the PPO-MD.
- Qualitative experiments, quantitative experiments, and ablation studies are designed to demonstrate the effectiveness and robustness of the proposed MLV-ODA and QLT-SLC method.

<sup>\*</sup>The corresponding author of this paper.

<sup>&</sup>lt;sup>1</sup>The authors are with the Institute of Robotics and Automatic Information System, College of Artificial Intelligence, Nankai University, Tianjin 300353, China, and also with the Tianjin Key Laboratory of Intelligent Robotics, Nankai University, Tianjin 300350, China (e-mail: 1120230216@mail.nankai.edu.cn; liujt@nankai.edu.cn).

This work is supported by fundation.



Fig. 1. Overview of the proposed system which mainly contains three parts: MLV-ODA method, QLT-SLC method and PPO-MD.

#### II. RELATED WORK

## A. Semantic Data Association

With the need for object construction in semantic SLAM, some object-level semantic SLAM algorithms have emerged, such as CubeSLAM [11] using cubes as object representations and QuadricSLAM [12] using quadrics. However, these methods did not address the crucial data association issue. Therefore, resolving the data association problem in object-level SLAM became a concern. Wu Y et al. [13] built on these two solutions and proposed the ensemble data association algorithm in EAO-SLAM, utilizing non-parametric tests, one-sample t-tests, and two-sample t-tests to indirectly associate objects through point cloud association. However, this algorithm does not leverage the pose and scale information of semantic objects, making it ineffective under conditions of missing or poor-quality point cloud information.

In this context, Tian R et al. [14], focusing on outdoor scenes, proposed an object data association algorithm that models data association as an assignment problem, constructing multiple association distances containing point cloud and object information and then obtaining the distance matrix by weighted summation, ultimately solving the assignment problem using the Hungarian algorithm [15]. Our previous work [10] continued this line of thought, improving it to adapt to indoor scenes, and proposed a joint data association algorithm. Although these two solutions partially addressed issues such as occlusion, lighting, and missed detections in complex scenes, as the problem scale increases, the space complexity of data association search and the time complexity of Hungarian algorithm solutions also increase, leading to reduced algorithm efficiency and performance degradation.

#### B. Semantic Loop Closure

Recently, many solutions propose constructing object-level semantic maps for scenes and then using the parameters of objects in the map and the layout between objects for scene representation, comparing differences between objects for loop detection and loop correction. Lin S et al. [16] proposed a loop closure method based on object construction and semantic graph matching, using voxels and cuboids to model objectlevel features in the environment and further representing the environment as a semantic graph with topological information. Based on this, an efficient graph matching method based on edit distance was proposed for robust location recognition. Finally, loop closure correction was performed through object alignment between semantic graphs. This method can effectively deal with significant viewpoint changes, but when the scene contains duplicate objects with similar topology, graph matching often leads to false alarms. Addressing this issue, Qian Z et al. [17] also proposed an approach called SmSLAM+LCD, integrated into the semantic SLAM system, combining high-level 3D semantic information and low-level feature information for accurate loop closure and effective drift reduction. Additionally, Yu J et al. [18] proposed SemanticLoop, modeling objects as TSDF and representing the environment as a 3D graph with semantics and topology, which corrects accumulated errors through aligned matching objects.

Since these methods have achieved very high precision and recall rates in the experiments, hence, we continue the basic ideas of these approaches, aiming to further improve the robustness of loop closure algorithms in complex scenes and their ability to handle significant viewpoint changes, as well as enhance the precision and recall of loop closure method.

#### **III. SYSTEM OVERVIEW**

Fig.1 illustrates the proposed framework, where the MLV-ODA method highlighted in the red box and the QLT-SLC method highlighted in the blue box, which represent the core algorithms of this paper. The PPO-MD in the green box is provided by our previous algorithm [10]. We have integrated the MLV-ODA method and the QLT-SLC method as two modules into the previous algorithm [10], creating a complete object-level semantic SLAM system. However, as the system has been extensively described in the previous paper [10], the other modules in the overall framework are omitted. This paper focuses on introducing the MLV-ODA method, the QLT-SLC method, and the PPO-MD.

**MLV-ODA method:** It involves 2D Frame IoU Verification, Label Posterior Probability Verification, 3D Quadric BackProj IoU Verification, and 3D Point BackProj Num Verification. The four hierarchical verification levels progressively narrow down the search space for data association, aiming to find a landmark association for each detection result as much as possible. At the same time, the association results are also provided to PPO-MD.

**QLT-SLC method:** It employs Frame and Quadric Landmarks Data Association, 3D Semantic Topological Graph Extraction and Covisibility Graph Calculation of Rotation Translation and Scale to obtain the loop closure candidates, which will be provided to PPO-MD for loop correction.

**PPO-MD:** It maintains KeyFrames of Camera Trajectory, Semi Dense Point Cloud Map, Sparse Point Cloud Map, and Quadric-Level Object Map. Simultaneously, it provides the required object and point cloud data for the MLV-ODA method and the QLT-SLC method and continuously updates its own database information based on the processing results of the MLV-ODA method and the QLT-SLC method.

#### IV. THE MLV-ODA METHOD

We will introduce the specific roles of each level in our proposed MLV-ODA method. There is a progressive relationship between levels, and each level will filter out the data that can be processed by that level. By adjusting the strategy and search space, we try our best to ensure that each detection can find the corresponding associated landmark.

## A. 2D-Level Verification For Micromotion Between Frames

In general, during the operation of SLAM, the system's frame rate is relatively high, and the scene change between adjacent frames is relatively small in the time sequence. Consequently, the movement amplitude of objects in the images is also relatively small. If the same object can be detected in both frames, the bounding box of the object in the two-dimensional pixel level of the image would also have a small movement range, resulting in a relatively large IoU. Therefore, we propose 2D Frame IoU Verification to preliminarily associate the current frame with the same object in the previous frame. Since data association has already been completed in the previous frame image, it is possible to indirectly associate the current frame detection results with the quadric landmarks associated with the previous frame. We define  $D_k^j$  as the *j*-th detection box in the current frame,  $D_{k-1}^{Q_i}$  as the detection

box associated with the quadric landmark  $Q_i$  in the previous frame, and  $IoU_{ij}$  is defined as follows:

$$IoU_{ij} = \frac{D_{k-1}^{Q_i} \cap D_k^j}{D_{k-1}^{Q_i} \cup D_k^j}$$
(1)

For  $D_k^j$ , we search for its maximum IoU value as follows:

$$max(IoU_{ij}) = max\{IoU_{0j}, \cdots, IoU_{ij}, \cdots, IoU_{nj}\}$$
(2)

Additionally, for robustness considerations, we implement category validation and threshold validation to further determine the success of data association. Association is considered successful only if condition  $IoU_{ij} > \delta_1$  and  $Label(D_k^j) = Label(Q_i)$  is satisfied.

#### B. Pro-Level Verification For False Positive Detection

Due to the presence of noise in the training dataset, the dataset's incompleteness, and the complexity of the testing environment, the object detection network is prone to false detections. Such false detections can render the 2D Frame IoU Verification ineffective, resulting in failed associations for some objects and subsequently leading to data association interruptions. To address the challenge, we propose the Label Posterior Probability Verification, as referenced in [19]. This method involves the probabilistic modeling of the quadric landmarks. Due to the discreteness and uncertainty of the categories, we utilize the Dirichlet process to model the categories of quadric landmarks. According to the bayesian probability model, the posterior probability that the detection box  $D_k^j$  belongs to the quadric landmark  $Q_i$  can be calculated as follows:

$$Posterior Pro_{ij} \propto Dirichlet Prior(Q_i) \cdot \\ Lable Likelihood(D_k^j) \cdot \\ Pos Likelihood(D_k^j, Q_i)$$
(3)

Through traversal, we find the maximum probability  $max(PosteriorPro_{ij})$  that satisfies the condition. If this maximum probability is less than the Dirichlet prior probability that it belongs to a new object, then we give more credibility to this prior probability, indicating that the current detection box should be associated with a new object, resulting in association failure. If this probability is greater than the prior probability of it belonging to a new object, we give more credence to this probability, indicating that the current detection has successfully associated with the object.

## C. 3D-Level Verification For Leak Detection and Continuity

Due to the limited scope of the previous two verification methods, which are based only on the detection box data of the previous frame and its associated quadric landmarks, they cannot effectively handle occlusions and lighting conditions that may cause missed detections in the object detection network.

Therefore, to ensure uninterrupted data association and stable long-term system operation, and also considering the issue of search time complexity, we employ a sliding window approach to appropriately expand the search space for quadric landmarks in the data association. Specifically, with the current frame as the reference, we select the nearest M keyframes in the time series and consider the collection of all quadric landmarks associated with them as the search space for the current data association process.

Firstly, for landmarks with existing quadric parameters, we propose the 3D Quadric BackProj IoU Verification. This method involves the back-projection of the quadric parameters of the landmark onto the image to form a projected bounding box. This bounding box is then used to calculate the IoU with the detection bounding box on the image. Let  $D_k^j$  denote the *j*-th detection box in the current frame,  $ProjD_k^{Q_i}$  represent the quadric landmark  $Q_i$  projected onto the detection box in the current frame. Some provide the section box in the current frame.

$$IoU_{ij} = \frac{ProjD_k^{Q_i} \cap D_k^j}{ProjD_k^{Q_i} \cup D_k^j}$$
(4)

The computation of the projected detection box  $D_k^{Q_i}$  is as follows:

$$D_k^{Q_i} = BBox(PQ_iP^T) \tag{5}$$

where  $P=K[R|t] \in \Re^{3\times 4}$  represents the projection matrix containing intrinsic and extrinsic parameters,  $Q_i \in \Re^{4\times 4}$ is a symmetric matrix with 9 degrees of freedom,  $PQ_iP^T$ represents the projection of the quadric into a conic, and  $BBox(\cdot)$  represents the bounding box fitting operator.

Similar to the approach in section E, we find the maximum value  $max(IoU_{ij})$  and then use the threshold  $\delta_2$  to further validate the data association.

Next, for landmarks without quadric parameters, we propose the 3D Point BackProj Num Verification. This involves projecting the associated map points of the landmarks back onto the image, generating projected feature points within different detection boxes, and then calculating the ratio of projected feature points falling into each detection box. We define  $D_k^j$  as the *j*-th detection box in the current frame,  $Point(Q_i)$  as all projected feature points of the quadric landmark  $Q_i$ ,  $Point(D_k^j)$  as the projected feature points falling into detection box  $D_k^j$ . Thus, the projected ratio  $Proportion_{ij}$  is defined as:

$$Proportion_{ij} = \frac{Point(D_k^j)}{Point(Q_i)} \tag{6}$$

Upon finding the maximum value  $max(Proportion_{ij})$ , we use the threshold  $\delta_3$  to further validate the data association.

# V. THE QLT-SLC METHOD

Algorithm 1 presents the core data flow of our proposed QLT-SLC method. The algorithm takes the current keyframe and the map database as input and outputs the paired loop closure candidates with the current frame and the best matching score. The processing steps are as follows: First, based on the number of co-observable objects and their ID differences, the algorithm performs an initial filtering of the candidate frames from the map database that meet the criteria, resulting in a set of loop closure candidates. Then, it initializes the optimal loop closure matching pair and the similarity threshold. Next, it traverses the set of loop closure candidates, extracting semantic topological graphs, including semantic nodes and semantic vectors, for each candidate frame and the current frame. Subsequently, it calculates the semantic similarity between the semantic topological graphs of the two frames based on their positions, rotations, and scales. This process is iterated

Algorithm 1: Core Data Process Of QLT-SLC Method **Input:** Current KeyFrame  $I_k$  and Map Database **Output:** Best Loop Closure  $(I_k, I_c^{best}, bestscore)$ 1 // filter candidate keyframes based on objs and ids 2  $\{I_c\} \leftarrow LoopMatchFilter(I_k, Th_{objs}, Th_{ids})$ 3  $(I_k, I_c^{best}, bestscore) \leftarrow O$ 4 for each  $I_c$  of  $\{I_c\}$  do 5 // extract semantic nodes and vectors for keyframes  $\{\vec{v}\}_k \leftarrow 3DSemTopoGraphExtract(I_k)$ 6  $\{\vec{v}\}_c \leftarrow 3DSemTopoGraphExtract(I_c)$ 7 // compute similarity socre for keyframes 8  $score \leftarrow CovGraphSimilarityCal(\{\vec{v}\}_k, \{\vec{v}\}_c)$ 9 // update best similarity score and best candidate 10 11  $bestscore \leftarrow bestscore > score?bestscore : socre$  $I_c^{best} \leftarrow I_c$ 12 13 end 14 if  $bestscore > Th_{score}$  then  $isTrueLoop \leftarrow ConsistencyCheck(I_k, I_c^{best})$ 15 if *isTrueLoop* then 16 // perform final loop correction 17  $LoopClosureCorrection(I_k, I_c^{best})$ 18 end 19 20 end

until the best similarity loop closure matching pair is found. Finally, the algorithm evaluates the similarity score threshold and checks for keyframe consistency. If a successful match is determined, the algorithm proceeds to perform loop closure correction based on the current keyframe and the loop closure candidate frame.

## A. Loop Closure Candidates Match Preprocessing

Through MLV-ODA, we have obtained association results between detections and landmarks. Before matching the candidate keyframes for the current keyframe, we first preliminarily filter whether there are enough co-observed landmarks and sufficient ID differences between the two keyframes. However, Directly comparing the differences between two keyframes would result in a high time complexity and relatively low efficiency. Therefore, we propose an efficient maintenance strategy, as shown in Fig.2. The first column represents the ObjectIndex, which maintains landmarks of different categories to quickly index the same object. The second column represents the KeyFrameQueue, which stores the queue of keyframes that observe the corresponding object. Whenever a keyframe is inserted into the map, each data association result in the keyframe is traversed. If the indexed object landmark exists in the ObjectIndex, the corresponding KeyFrameQueue is updated with the current keyframe. Otherwise, a new ObjectIndex and KeyFrameQueue are created for maintenance.

With this maintenance strategy, we can quickly and efficiently find a candidate keyframe set for the current keyframe, where the number of co-observed object landmarks is greater than or equal to  $Th_{objs}$  and the ID difference between the keyframes is greater than or equal to  $Th_{ids}$ .



Fig. 2. Efficient maintenance strategy based on ObjectIndex and KeyFrame-Queue for covisibility information.

## B. 3D Semantic Topological Graph Extraction

After obtaining the candidate keyframe set, we proceed to extract the 3D semantic topological graph for each keyframe, which includes 3D semantic nodes and 3D semantic vectors. The specific operations are as follows:

Let  $I_i$  represent the current keyframe,  $I_j$  represent the candidate keyframe, and  $\{(Q_p^i, Q_q^j)\}$  represent the set of quadriclevel semantic nodes associated with  $I_i$  and  $I_j$ . Let  $C_p^j = (X_p^j, Y_p^j, Z_p^j)$  be the center point coordinates of semantic node  $Q_p^i$  in the  $I_i$  coordinate system, and  $C_q^j = (X_q^j, Y_q^j, Z_q^j)$  be the center point coordinates of semantic node  $Q_q^i$  in the  $I_j$  coordinate system. Here, the topological structure between quadric landmarks is defined as the semantic vector formed by the lines connecting the center point coordinates of each pair of semantic nodes. For example,  $\vec{v} = \overline{C_1^k C_2^k}$  represent the semantic vector between the 1st and 2nd semantic nodes in the k-th frame.

The semantic vector between the 1st and 2nd semantic nodes in the current keyframe  $I_i$  is:

$$\vec{v}_i = (v_{ix}, v_{iy}, v_{iz})^T = \overline{C_1^i C_2^i}$$
 (7)

The semantic vector between the 1st and 2nd associated semantic nodes in the candidate keyframe  $I_i$  is:

$$\vec{v}_j = (v_{jx}, v_{jy}, v_{jz})^T = \overline{C_1^j C_2^j}$$
(8)

According to the properties of vectors in space, the semantic vectors  $v_1$  and  $v_2$  differ by a rotation R and scale s, that is:

$$\vec{v}_j = R(s\vec{v}_i) \tag{9}$$

Here,  $R \in \Re^{3 \times 3}$  represents the rotation matrix between the two vectors, and s is a scalar representing the scale factor between the two vectors. We can solve for R and s using the following Rodrigues's Formula:

$$R = \cos\theta I + (1 - \cos\theta)\vec{\omega}\vec{\omega}^T + \sin\theta\vec{\omega}^\wedge \tag{10}$$

$$s = \frac{|\vec{v}_j|}{|\vec{v}_i|} \tag{11}$$

Where  $\theta$  are the rotation angle, and  $\vec{\omega} = (\omega_x, \omega_y, \omega_z)^T$  is the rotation axis. The calculation of the rotation angles and rotation axis is as follows:

$$\theta = \arccos \frac{\vec{v}_i \vec{v}_j}{|\vec{v}_i| |\vec{v}_j|} \tag{12}$$

$$\vec{\omega} = \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} = \begin{pmatrix} v_{iy}v_{jz} - v_{iz}v_{jy} \\ v_{iz}v_{jx} - v_{ix}v_{jz} \\ v_{ix}v_{jy} - v_{iy}v_{jx} \end{pmatrix}$$
(13)

In addition to considering the rotation and scale between the semantic vectors, we also consider the gap  $\vec{t}$  between the starting points of the semantic vectors. Considering the starting point of the semantic vector  $\vec{v}_i$  as  $C_1^i$  and the starting point of  $\vec{v}_j$  as  $C_1^j$ , the corresponding translation  $\vec{t}$  is:

$$\vec{t} = (X_1^j - X_1^i, Y_1^j - Y_1^i, Z_1^j - Z_1^i)^T$$
(14)

Thus, we have established the three indicators, R, s, and t, to measure the similarity between two semantic vectors. When the two semantic vectors coincide, the three indicators are  $R = I_{3\times3}$ , s = 1, and  $t = (0, 0, 0)^T$ , where I is the unit matrix.

## C. Covisibility Graph Similarity Calculation

Based on the definitions and transformation relationships provided for the semantic vectors, we can proceed with the similarity calculation for the covisibility graph in the 3D semantic topological graph. The specific steps are as follows:

Initialize the spatial semantic similarity Score = 0.

Traverse the semantic node set of the current keyframe  $I_i$  and the candidate keyframe  $I_j$ .

Calculate the rotation, scale, and translation indicators  $\{R, s, \vec{t}\}$  between the semantic vectors formed by each pair of semantic nodes in the current keyframe and the corresponding two semantic nodes in the candidate keyframe.

Calculate the topological structure similarity score between the two semantic vectors:

$$Score_{topology} = e^{-\frac{1}{2}|1-s||R|| - ||\vec{t}||}$$
 (15)

Update the semantic similarity score:

$$Score_{semantic} = Score_{semantic} + \rho Score_{topology}$$
 (16)

Update the score of the spatial semantic similarity between the topological graph:

$$Score = Score + Score_{semantic}$$
 (17)

After traversal is completed, output the final spatial semantic similarity score between the two topological graphs.

We use formula (11) to calculate the similarity between two semantic vectors in the topological graph. Its characteristic is that as  $s ||R|| + ||\vec{t}||$  approaches 1, the function value increases faster and faster, so it is easier to process than linear The function is more strict because we hope that the three indicators between two semantic vectors will have a higher similarity score only when they are close to  $\{||R|| = 1, s = 1, ||\vec{t}|| = 0\}$ .

In addition, this paper presents  $\rho Score_{topology}$  when calculating the semantic similarity score using formula (16), where  $\rho$  represents the quality score of the quadric landmark corresponding to the starting semantic node of the semantic vector in the topological graph (which is obtained from the EQI algorithm in the previous work [10]). The purpose of using the value  $\rho$  is that when the parameter quality score of the constructed quadric landmark is relatively small, it means that the confidence of the quadric landmark is relatively low, so its impact on loop closure detection should be smaller.





Fig. 4. Ablation study performance on semantic object construction in the TUM dataset fr2-desk sequence.

## VI. EXPERIMENTS AND EVALUATION

We evaluated our proposed method using the TUM RGB-D benchmark [20], which is a commonly used benchmark for evaluating the performance of visual SLAM algorithms. All experiments were conducted on our local device, with the following specifications: AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz, 16.0 GB RAM, Nvidia RTX 3060 6GB.

#### A. The Performance of MLV-ODA Method

Data association performance cannot be measured directly, but it impacts the semantic objects construction performance. Therefore, we indirectly validate the effectiveness and robustness of the proposed MLV-ODA method using the performance of semantic objects construction. We introduce indicators precision and recall to measure the accuracy and completeness of semantic object construction respectively. We judge that only when the category of the object is consistent with the real scene and there are no other constructed objects in the vicinity of the object, it can be considered a true positive.

Based on the performance shown in both the Table I and the Fig.3, our proposed method exhibits favorable results. Qualitatively, our method accurately constructs semantic objects in the environment, including poses and labels, with minimal occurrences of erroneous construction, missing construction, or redundant construction. Quantitatively, our method achieves a high level of performance, with an average precision of 97.49%, an average recall of 85.88%, and an average F-value of 91.31%.

To further validate the effectiveness of the proposed hierarchical approach in the MLV-ODA method, we conducted ablation experiments by isolating different levels within the MLV-ODA method. Specifically, we extracted three data association methods: 2D-ODA, which exclusively uses 2D-level verification, 3D-ODA, which solely relies on 3D-level verification, and Pro-ODA, which employs only probabilistic-level verification. We then compared these methods with our MLV-ODA algorithm.

As observed from the performance in the Table III and the Fig.4, our algorithm achieves the best Precision and Recall, highlighting the effectiveness of the proposed multilevel verification approach. The 2D-ODA method exhibits the poorest performance mainly due to the instability of the object detection network, resulting in a high frequency of interruptions in data association due to numerous false positives and false negatives. Although the 3D-ODA method achieves the same Precision as MLV-ODA, it suffers from lower recall. This is because although 3D-ODA can handle some cases of missed detections and occlusions, insufficient data on object association due to observation limitations leads to failed semantic objects construction. However, the Pro-

 TABLE I

 PRECISION AND RECALL EXPERIMENT OF OUR MLV-ODA METHOD ON FOUR SEQUENCES OF TUM DATASET

Sequence	fr1-room	fr2-desk	fr2-dishes	fr3-longoffice	Average
Num of Sequence	1352	2893	1706	2488	2109.75
Num of Constructed Object	17	14	5	24	15
Num of RealExist Object	19	16	6	26	16.75
True Positive	16	14	5	23	14.5
Precision	94.12%	100.00%	100.00%	95.83%	97.49%
Recall	84.21%	87.50%	83.33%	88.46%	85.88%
F-Measure	88.89%	93.33%	90.91%	92.00%	91.31%

TABLE II

PRECISION AND RECALL ABLATION STUDY ON FR2-DESK SEQUENCE OF TUM DATASET

Method	2D-ODA	3D-ODA	Pro-ODA	Our MLV-ODA
Num of Constructed Object	125	9	17	14
Num of RealExist Object	16	16	16	16
True Positive	1	9	13	14
Precision	0.80%	100.00%	76.47%	100.00%
Recall	6.25%	56.25%	81.25%	87.50%
F-Measure	1.42%	72.00%	78.79%	93.33%

TABLE III

Comparison of average time consumption between JDA and MLV-ODA at various stages of operation(MS)

Sequenece	Metric	JDA	MLV-ODA
	10	0.0758	0.0454
fr2-desk	100	0.1867	0.0420
	1000	2.8692	0.0418
	10	0.0128	0.0812
fr3-longoffice	100	0.1907	0.0732
2	1000	2.9550	0.0450

ODA method performs comparably to our proposed MLV-ODA, primarily due to the integration of semantic category and object position information. Nevertheless, the assumed probabilistic model may not be universally applicable to all objects, resulting in some unsuccessful associations.

In addition, we also tested the efficiency of our MLV-ODA method compared with the JDA method using the Hungarian algorithm that we previously proposed in [10]. In order to make the test intuitive, we recorded the method running time in three stages (10 frames, 100 frames and 1000 frames). The data in Table VI-A can be seen that the time consumption of our method in each stage is not much different and is very stable. However, due to the time complexity of the Hungarian algorithm, the time consumption of the JDA method will surge as the number of frames increases.

## B. The Performance of QLT-SLC Method

We selected two recent outstanding approaches, [17] and [18], as the comparative methods for our QLT-SLC method in this paper. However, their criteria for selecting reference loop closures are different. The criteria proposed in [17] for selecting reference loop closure candidates are as follows: a position difference of within 1m, an angle difference of within 53°, and an ID difference of over 1000. The criteria proposed in [18] are based on the condition that the number of commonly observed objects between two frames is greater than 2, and the ID difference is over 500. However, in our practice, using the criteria proposed in [17], we do not can obtain a similar number of reference loop closures. Therefore, for the fairness, after our experiment, we use the following selection criteria: a position difference of within 3m, an angle difference of within 80°, and an ID difference of over 1000, which actually adds difficulty to our method. The Fig.5 shows the statistical graphs of selected reference loops in the fr2-desk and fr3-longoffice sequences of the TUM dataset for our method.



Fig. 5. Reference loop closure statistics under condition: position difference within 3m, angle difference within  $80^\circ$  and ID difference above 1000

The Table IV presents the comparative experiments of our method and three other methods, including traditional loop closure methods [4] and [5], as well as semantic loop closure methods [17] and [18]. Among them, Ours-C1 represents the result of our method under the matching configuration of [17], and Ours-C2 represents the result of our method under the matching configuration of [18]. The main comparison metrics are common in loop closure, namely precision and recall. From the table, we observe that our QLT-SLC method, under both configurations, outperforms the original algorithms. Additionally, our method not only detects more loop closures but also exhibits robustness under significant viewpoint changes.

The reason why our method performs well is that on the one hand, thanks to the MLV-ODA method, the data association between the current frame and the map is very robust. We can find more object information for the current frame, so that even under large viewing angle changes, we can find a lot of TP loop closures. On the other hand, the similarity calculation we proposed combines spatial information and semantic information, as well as loop closure consistency check, which can avoid FP loop closures caused by similar image scenes.

## C. The Performance of Localization

We use the ATE metric to do localization accuracy comparison experiments. Data for [4], [5], and our method were generated on our local device, while the data for the three approaches [16], [17] and [18] were obtained directly from the respective papers as they were not publicly available.

TABLE IV						
COMPARATIVE EXPERIMENT ON PRECISION	AND RECALL OF LOOP CLOSURE (	ON TWO SEQUENCES OF TUM DATASET				

Sequence	Metrics	ORB-SLAM2	ORB-SLAM3	[17]	[18]	Ours-C1	Ours-C2
fr2-desk	KeyFrame	193	261	/	/	192	/
	ReferenceLoop	1697	2774	2122	/	2079	/
	RealLoop	176	361	23	44	45	55
	True Positive	33	66	23	44	45	55
	Precision	18.75%	18.28%	100.00%	100.00%	100.00%	100.00%
	Recall	1.94%	2.38%	1.08%	/	2.16%	/
fr-longOffice	KeyFrame	239	293	/	/	274	/
	ReferenceLoop	1773	2654	2429	/	2188	/
	RealLoop	209	504	6	76	25	85
	True Positive	16	53	6	76	25	85
	Precision	7.66%	10.52%	100.00%	100.00%	100.00%	100.00%
	Recall	0.90%	2.00%	0.25%	/	1.14%	/
Confi	guration	3m,80°,1000	3m,80°,1000	1m,53°,1000	2,500	3m,80°,1000	2,500

		_	
ΤA	BL	Æ	v

Comparative experiment of localization accuracy after loop closure on TUM dataset(M)  $% \mathcal{M}$ 

Sequence	Metrics	ORB-SLAM2	ORB-SLAM3	[16]	[17]	[18]	Ours
	RMSE	0.008	0.018	0.014	0.008	0.042	0.0078
fr2-desk	MEAN	0.008	0.017	/	/	/	0.0072
	MEDIAN	0.007	0.015	/	/	/	0.0068
	RMSE	0.012	0.012	0.016	0.009	0.015	0.0087
fr3-longoffice	MEAN	0.011	0.011	/	/	/	0.0079
	MEDIAN	0.009	0.010	/	/	/	0.0075

As shown in the Table V, since our method can detect more accurate loop closure, therefore, our method exhibits the best performance. Even compared with the latest approach [18], achieving competitive accuracy results.

## VII. CONCLUSION

In this paper, to address the issue of inaccurate data association under scenarios such as false positives, false negatives, and occlusions, we propose the MLV-ODA method, which reduces the time and space complexity of solving the data association problem, thereby improving efficiency while ensuring accuracy. To tackle the problem of non-robust loop closure detection in scenarios involving significant viewpoint changes or similar scenes, we introduce the QLT-SLC method, which outperforms existing state-of-the-art methods, enhancing both the accuracy and recall of the loop closure detection process. However, our proposed method still has some limitations, such as not considering dynamic scenes and not involving semantic objects in loop correction. Therefore, in future work, we aim to address these limitations and further enhance the robustness of the proposed method.

#### REFERENCES

- [1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.
- [2] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in 2011 International conference on computer vision. Ieee, 2011, pp. 2564–2571.
- [3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [4] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions* on robotics, vol. 33, no. 5, pp. 1255–1262, 2017.
- [5] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
- [6] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

- [7] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.
- [8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779– 788.
- [9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
- [10] Z. Cao, Y. Zhang, R. Tian, R. Ma, X. Hu, S. Coleman, and D. Kerr, "Object-aware slam based on efficient quadric initialization and joint data association," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9802–9809, 2022.
- [11] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.
- [12] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2018.
- [13] Y. Wu, Y. Zhang, D. Zhu, Y. Feng, S. Coleman, and D. Kerr, "Eao-slam: Monocular semi-dense object slam based on ensemble data association," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 4966–4973.
- [14] R. Tian, Y. Zhang, Y. Feng, L. Yang, Z. Cao, S. Coleman, and D. Kerr, "Accurate and robust object slam with 3d quadric landmark reconstruction in outdoors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1534–1541, 2021.
- [15] H. W. Kuhn, "The hungarian method for the assignment problem," Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.
- [16] S. Lin, J. Wang, M. Xu, H. Zhao, and Z. Chen, "Topology aware object-level semantic mapping towards more robust loop closure," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7041–7048, 2021.
- [17] Z. Qian, J. Fu, and J. Xiao, "Towards accurate loop closure detection in semantic slam with 3d semantic covisibility graphs," *IEEE Robotics* and Automation Letters, vol. 7, no. 2, pp. 2455–2462, 2022.
- [18] J. Yu and S. Shen, "Semanticloop: loop closure with 3d semantic graph matching," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 568–575, 2022.
- [19] B. Mu, S.-Y. Liu, L. Paull, J. Leonard, and J. P. How, "Slam with objects using a nonparametric pose graph," in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2016, pp. 4602–4609.
- [20] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in 2012 IEEE/RSJ international conference on intelligent robots and systems. IEEE, 2012, pp. 573–580.